Mining of Frequent OptimisticEstimations by Using Measured Techniques

[1]M. Lavanya, [2]Dr. M. Usha Rani
[1]Dept. of Master of Computer Applications Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati, AP, India
[2]Dept. of Computer Science, Sri Padmavati Mahila Viswavidyalayam, (SPMVV Woman's' University), Tirupati, AP, India

Abstract
In recent years the sizes of databases has increased rapidly. This has led toa growing interest in the development of tools capable in the automatic extractionof knowledge from data. The term Data Mining, or Knowledge Discovery inDatabases, has been adopted for a field of research dealing with the automaticdiscovery of implicit information or knowledge within databases.Several efficient algorithms have been proposed for finding frequentitemsets and the association rules are derived from the frequent itemsets, such as theApriori algorithm. These Apriori-like algorithms suffer from the coststo handle a huge number of candidate sets and scan the database repeatedly. A frequent pattern tree (FP-tree) structure for storing compressed and criticalinformation about frequent patterns is developed for finding the complete set of frequent itemsets. But this approachavoids the costly generation of a large number of candidate sets and repeated databasescans, which is regarded as the most efficient strategy for mining frequent itemsets.Finding of infrequent items gives the positive feed back to the Production Manager.
In this paper, we are finding frequent and infrequent itemsets by taking opinions of different customers by using Dissimilarity Matrix between frequent and infrequent items and also by using Binary Variable technique. We also exclusively use AND Gate Logic function for finding opinions of frequent and infrequent items. After finding frequent and infrequent items the apply Classification Based on Associations (CBA) on them to have better classification.

Keywords
Knowledge Discovery, Frequent Items, Infrequent Items, Similarity, Dissimilarity

I. Introduction

Finding of associations between items from a large database of businessdata, has been a latest topic within the area of data mining [1-2]. The effective management of business is significantly dependent on the quality of its decision making. These are useful in market basket analysis and catalog design .It is therefore important to analyze past transaction data to discover customer purchasing behavior and improve the quality of business decision.There are several techniques are available to find frequent items. The strategies for mining frequent itemsets, includes Apriori [1], and FPgrowth [3]. To support the above analysis, collect the transaction items based on requirement and store it in a database. The major work of mining frequent itemset is to find all itemsets that satisfy a certain user-specified minimumsupport. Each such item set is referred to as large item set. The rest of the paper is organized as follows: Section II, presents the existing work Section III, proposes the efficient data mining algorithms for finding opinions of frequent items.
In day-day activities,world data is being collected everywhere, and everyone is eager in extracting knowledge out of it and use it to improve their performance. Everyone has different motive for example, business people would need to improve their business and make profit out of it. Physicians may aim at discovering knowledge out of medical records to prevent disease. Network administrators would need to keep their network secure hence they would aim at detecting anomalies to prevent intrusion. It is impractical to manually analyze the data and extract knowledge as the volume of data is very high. Hence, we aim at finding patterns to discover knowledge from the raw data which is called data mining.
There are different definitions floating around for data mining and there are various techniques to find the patterns from the raw data. Two of such techniques to extract pattern from the raw data are Decision Trees and Neural Networks. In this paper, we will compare, contrast and evaluate J48 implementation of C4.5 Decision Tree algorithm and Multilayer Perception which is Neural Network based algorithm. Finally, we conclude this paper and present directions for future research in Section IV.

II. Existing Algorithms In Association Rules Mining

Association rule mining as introduced in [1], searches for relationship between items ina data set. It finds association, correlation, or casual structures among set of items or objects intransaction databases, relational databases and other information repositories. To mine anassociation rule, database of transaction is needed. And each transaction is list of items. Then apply mining algorithm to find the association rule.
Finding frequent itemsets plays an impotent role in the field of data mining. Frequentitem set are essential for many data mining problems like discovery of association rule correlation [4-5] and sequential pattern [6-7]. As defined in [2], the problem is stated as follows.
Let I = {x1,….xn} be a set of items. An itemset X is a subset of items, i.e. X □□I. Atransaction T = (itd, X) is a 2-tuple, where tid is a transaction-id and X an itemset. A transactionT=(tid, X) is said to be contain itemset Y if and only if Y is a subset of X. A transactiondatabase D is a set of transactions. The number of transactions in D containing itemset X iscalled the support of X. Given a transaction database D and a support threshold min_sup, anitemset X will be called as frequent pattern if and only if sup(X)_min_sup.

A. The Apriori Algorithm

Apriori algorithm finds all frequent itemsets. The first pass of the algorithm simply counts item occurrences to determine the large 1-itemsets. A subsequent pass, say pass k, consists of two phases. First, the large itemsets Lk-1 found in the (k-1)th pass are used to generate the candidate itemsets Ck, using the Apriori candidate generation function (apriori-gen). Next, the database is scanned and the support of candidates in Ck is counted. For fast counting, an efficient determination if the candidates in Ck that are contained in a giventransaction t is needed. A hash-tree data

structure [8] is used for this purpose.

B. The FP-Growth Algorithm

The main bottleneck of the Apriori-like methods is at the candidate set generation and test. This problem was dealt with by introducing a novel, compact data structure, called frequent pattern tree, or FP-tree then based on this structure an FP-tree-based pattern fragment growth method was developed, FP-growth.

This approach avoids the costly generation of a large number of candidate sets and repeated database scans, which is regarded as the most efficient strategy for mining frequent itemsets. The definition, according to [9] is as follows.

A frequent pattern tree (FP- Tree) is a tree structure defined below.

1. It consists of one root labeled as "root", a set of item prefix sub-trees as thechildren of the root, and a frequent-item header table.
2. Each node in the item prefix sub-tree consists of three fields: item-name, count,and node-link, where item-name registers which item this node represents, countregisters the number of transactions represented by the portion of the pathreaching this node, and node-link links to the next node in the FP-tree carryingthe same item-name, or null if there is none.
3. Each entry in the frequent-item header table consists of two fields, (1) item-nameand (2) head of node-link, which points to the first node in the FP-tree carryingthe item-name.

A fuzzy classification rule is a fuzzy if–then rule whose consequent part is a class label. Since the comprehensibility of fuzzy rules by human users is a criterion in designing a fuzzy rule-based system (Ishibuchi et al., 1999), fuzzy classification rules with linguistic interpretations must be taken into account. To cope with this problem, we consider both quantitative and categorical attributes, which are used to describe each sample data, as linguistic variables. Then, each linguistic variable can be partitioned by its linguistic values represented by fuzzy numbers with triangular membership functions. Simple fuzzy grids or grid partitions (Ishibuchi et al., 1999; Jang and Sun, 1995) in feature space resulting from the fuzzy partition are thus obtained.

III. Proposed Techniques

Section II, gives the information aboutonly for finding frequent items means which are frequently purchased by the customer. Generally in real time applications most of the people collect more information about the products before they are purchasing particular product. The finding of product features using existing mining algorithms is a difficult task. All the existing techniques concentrate on finding frequent or infrequent items only. No such existing algorithms are available to find positive opinions. But so many web sites allow the users to give their opinions about the product while they are purchasing or after using the product. Collecting either positive or negative opinions using existing algorithms is a difficult task. But our proposed algorithm can find the positive opinions of frequent items about the products. Before discussing about proposed technique let us first define some definitions:

Symmetric and Asymmetric Binary Variables

A binary variable is Symmetric if both of its states are equally valuable and carry the same weight otherwise it is asymmetric.

Dissimilarity between binary Variables

One approach involves computing the dissimilarity between all binary variables involves computing dissimilarity matrix. Dissimilarity between two binary variables can be calculated by using the following formula.

$d(i,j)= r+s/(q+r+s)$

where q is the number of variables that equal to 1for both objects i and j, r is the number of variables is equal to 1 for object i but that are 0 for object j and s is the number of variables is equal to 1 for object j but that are 0 for object i.Suppose that a Product item set or product relation list table (Table 2) contains the attributes carry, price, noise, vibration, durability, installation.Where Product name represents object identifier and the attributes specified in Table 2 all are Asymmetric attributes. For Asymmetric attributes we can set two values either 0 or 1.Carry related to implicit attribute weight will give two values

Table 1: Notational descriptions of Relational table

Easy	E	
Difficult	D	
Price returns	Reasonable	R
	More	M
Noise	Less	L
	More	M
Vibration	Less	L
	More	M
Durability	Long	L
	Small	S
Installation	Easy	E
	Difficult	D

A. Proposed Algorithm

Algorithm1: (Using Dissimilarity Matrix)

1. Construct relation table of the products and their attributes
2. Find the dissimilarity between each product and their attributes
3. Compare Dissimilarity of each product with minimum support set for the product item set
4. If (Dissimilarity of each product >minimum support) then
4.1 the given item is infrequent
Else
4.2 the given item is frequent
end if
5. Repeat step4 for all products in the product list or product item set

Table 2: Description of Relational Table of Products Using Binary Attributes

Product name\attribute	Carry	Price	Noise	Vibration	Durability	Installation
Washing Machine	D	R	L	M	L	E
Inverter	E	M	L	L	S	D
Cell phone	E	R	M	M	M	E
Air Conditioner	E	M	L	L	S	D
Refrigerator	D	R	M	L	M	E

Algorithm 2: (Using AND Gate Logic)

Consider two variables A and B. A represents Product list and B represents Cluster. Maintain all products which are produced to the customers are in A and B contains all positive attributes related to the products.

Example: Pen drive is easy to carry

Here carry is an attribute and easy is a positive attribute to the above sentence .For the same sentence Difficult to carry is the negative attribute.

Table 3: General function of AND Logic Gate

A	B	C
T	T	T
T	F	F
F	T	F
F	F	F

The Logic is as follows:

The algorithm consists of the following steps

case 1

If product belongs to Product List and if opinion is in the cluster

Then

Find the dissimilarity of the product

case 2

Else if the product belongs to product list but opinion is not in the cluster

Find the support of the product

case 3

Else if product does not belongs to product list and and opinion is in cluster

- No need to calculate dissimilarity between variables
- Apply the table 3 on each product in the product list and store the result of each case in a separate table
- Find the support of each table
- The result of case1 is for positive opinions
- Result of case 2 is for negative opinions

Both these two algorithms are new techniques to calculate positive opinions about the products.

The problem in Algorithm 1 is

- Setting of minimum support, it depends on Domain Expert. If it is high, rarenumber of associations between items gets. If it is low,more number of frequent items will be occurred.
- It calculates both frequent and infrequent items. So time taken to calculate frequent items will be high because it scans entire database.

But Algorithm 2 calculates and considers only positive opinions.

Algorithm III

Finding fuzzy classification rules based on the frequent items mined [10]. First provide training samples, where the size is determined by K. select maximum number of iterations for giving the rules of changing to deal with new situations by providing size and weights for items providing mutation probability by giving maximum numbers of items to be purchased for above algorithm which follows in two steps .

Phase I: determine frequent fuzzy grids;

Phase II: produce fuzzy classification rules

Method of implementing algorithm is as follows:

Step1. Initialization

Step 2. Perform the simple fuzzy partition

Step 3. Scan the training samples, and construct FGTTFS

Step 4. Compute the fitness

Step 5. Generate frequent fuzzy grids

Step 6. Generate fuzzy classification rules

Step 7. Reduce redundant rules

Step 8. Employ adaptive rules to adjust fuzzy confidences (Nozaki et al., 1996)

Step 9. Selection

Step 10. Crossover

Step 11. Mutation

Step 12. Elitist strategy

Step 13. Termination test

IV. Conclusion

The Proposed algorithm for discovering opinions of frequent item sets based on dissimilarity matrix using binary variable is a new method and is found efficient when compared to Apriori and FP-tree. As such, we are stillworking on it with the aim of extending the application of this algorithm to various kinds of databases. And apply classification algorithm to get the refined items to be mined.

References

[1] R. Agrawal, T. Imielinski, A. Swami,"Mining association rules between sets of items inlarge databases", Proc. 1993 ACM-SIGMOD Int.Conf: Management of Data, 1993, pp. 207-2 16.

[2] R. Agrawal, R. Srikant,"Fast algorithms for mining association rules", Proc. 20th Int. Conf:Very Large Data Bases, 1994, pp. 487-499.

[3] J. Han, 1. Pei, Y. Yin,"Mining Frequent Patterns without CandidateGeneration", in Proc. of the 2000 ACM SIGMOD InternationalConference on Management of Data, 2000.

[4] C.K.S. Leung, L.V.S. Lakshmanan, R.T. Ng.,"Efficient dynamic mining ofconstrained using FP-trees", SIGKDD Explorations, 4(1), pp. 40-49, 2002.

[5] J.Pei, J. Han, R.Mao.,"CLOSET: an efficient algorithm for mining frequent closed itemsets. In Proc. DMKD, pp. 21-30, 2000.

[6] Mannila H., Toivonen H.,"Discovering frequent episodes in sequences", In Proc. 1stInternational Conference on Knowledge Discovery and Data Mining (KDD), pp. 210-215, 1995.
[7] MannilaH, Toivonen H.,"Discovering generalized episodes using minimaloccurrences", In Proc. 2nd International Conference on Knowledge Discovery and Data Mining (KDD), pp. 146-151, 1996.
[8] R. Agrawal, R. Srikant,"Fast algorithms for mining association rules in largedatabases", Proc. of 20th Int'l conf. on VLDB, pp. 487-499, 1994.
[9] J. Han, J. Pei, Y. Yin,"Mining Frequent Patterns without CandidateGeneration", Proc. of ACM-SIGMOD, 2000.
[10] Yi-Chung Hu, Ruey-Shun Chen, Gwo-Hshiung Tzeng, "Finding fuzzy classification rules using data mining techniques", Institute of Information Management, National Chiao Tung University, Hsinchu 300, Taiwan, ROC.

M. Lavanya received her Bachelor's degree in Sciences (Computer Science) from S.V.University, Tirupathi, in 2000. Master's degree in Computer Applications from S.V.University, in 2003. She is working as Assistant Professor in the Department of Master of Computer Applications at Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati. She is pursuing her Ph.D. in Computer Science in the area of Data Warehousing and Data Mining. She is in teaching since 2003. She presented many papers at National and Internal Conferences and published articles in National & International journals. Her research interests include web mining, information retrieval systems.

Dr. M. UshaRani is an Associate Professor in the Department of Computer Science and HOD for MCA, Sri Padmavati Mahila Viswavidyalayam (SPMVV Woman's' University), Tirupati. She did her Ph.D. in Computer Science in the area of Artificial Intelligence and Expert Systems. She is in teaching since 1992. She presented many papers at National and Internal Conferences and published articles in national & international journals. She also has written 4 books like Data Mining - Applications: Opportunities and Challenges, Superficial Overview of Data Mining Tools, Data Warehousing & Data Mining and Intelligent Systems & Communications. She is guiding M.Phil. and Ph.D. in the areas like Artificial Intelligence, Data Warehousing and Data Mining, Computer Networks and Network Security etc.